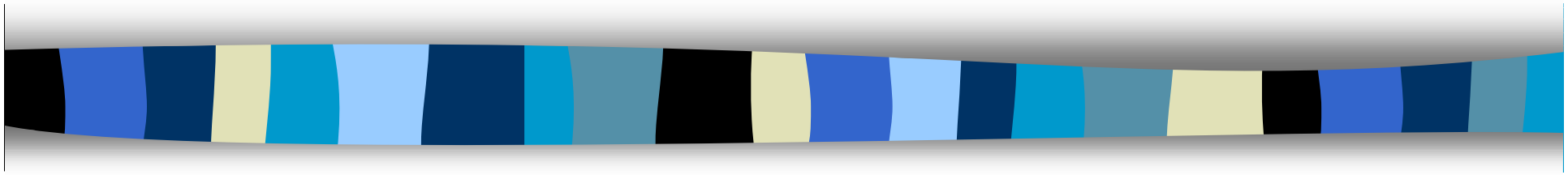
Klasifikace hudebních stylů



Martin Šimonovský

(mys007@seznam.cz)

Rozpoznávání hudby

- úloha z oblasti DSP

- klasifikace dle hudebních stylů <<<
  - zachycení obecných charakteristik, generalizace
  - domněnka: úloha vhodná pro NS
- rozpoznání skladby na základě ukázky
  - vypočtení robustního přesného otisku vzorků každé skladby
  - oblast rozvoje (komerční nasazení)

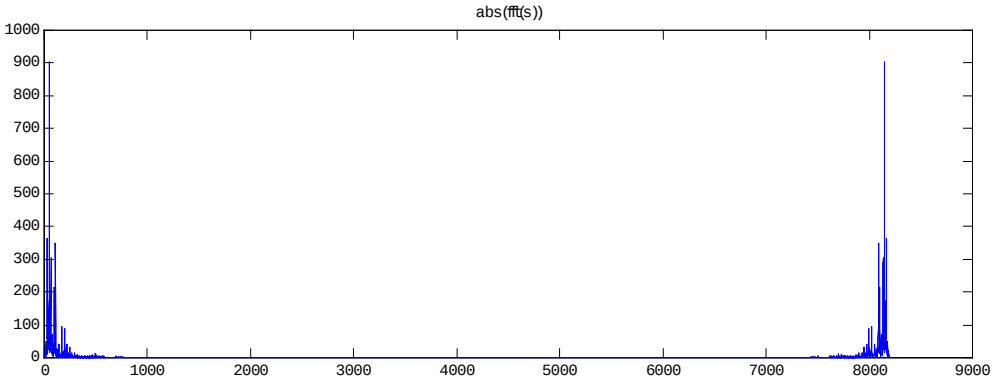
- podobná úloha: rozpoznávání mluvčího/řeči
  - existují relativně úspěšné algoritmy a charakteristiky

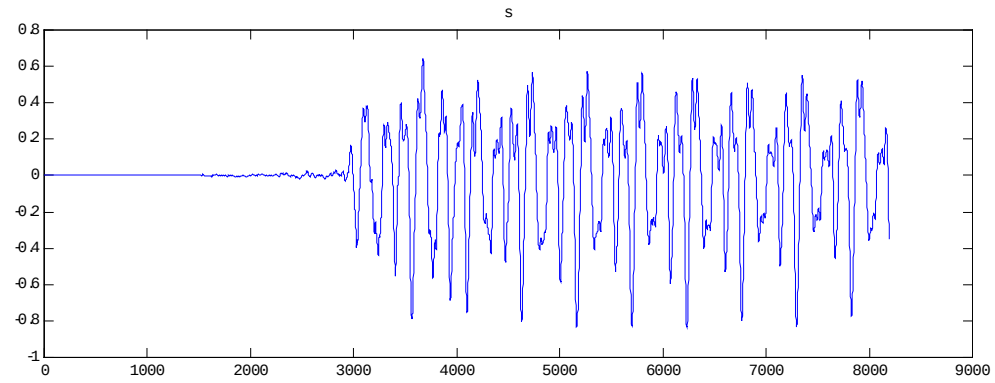
Krok stranou: lidský sluch

- tón - síla, výška, zabarvení
- frekvence
  - rozsah vnímání: 20Hz - 20kHz
  - stupnice je logaritmická:
    - tj. 50Hz-100Hz a 10kHz-20kHz obsahují zhruba stejně informací
  - oktáva = dvojnásobek frekvence
  - vzdálenost tónů: *2^(1/12)
  - nejdůležitější složka analýzy zvuku
- amplituda
  - opět logaritmická stupnice: dB
  - v rámci jedné skladby se výrazně nemění -> používám lineární stupnici
- fázi dvou vln lidské ucho nerozlišuje

Spektrální analýza signálu

- dvě reprezentace signálu v daném časovém intervalu
  - časová doména: y = f(vzorek)
    - není vhodná
  - frekvenční doména: y = f(frekvence)

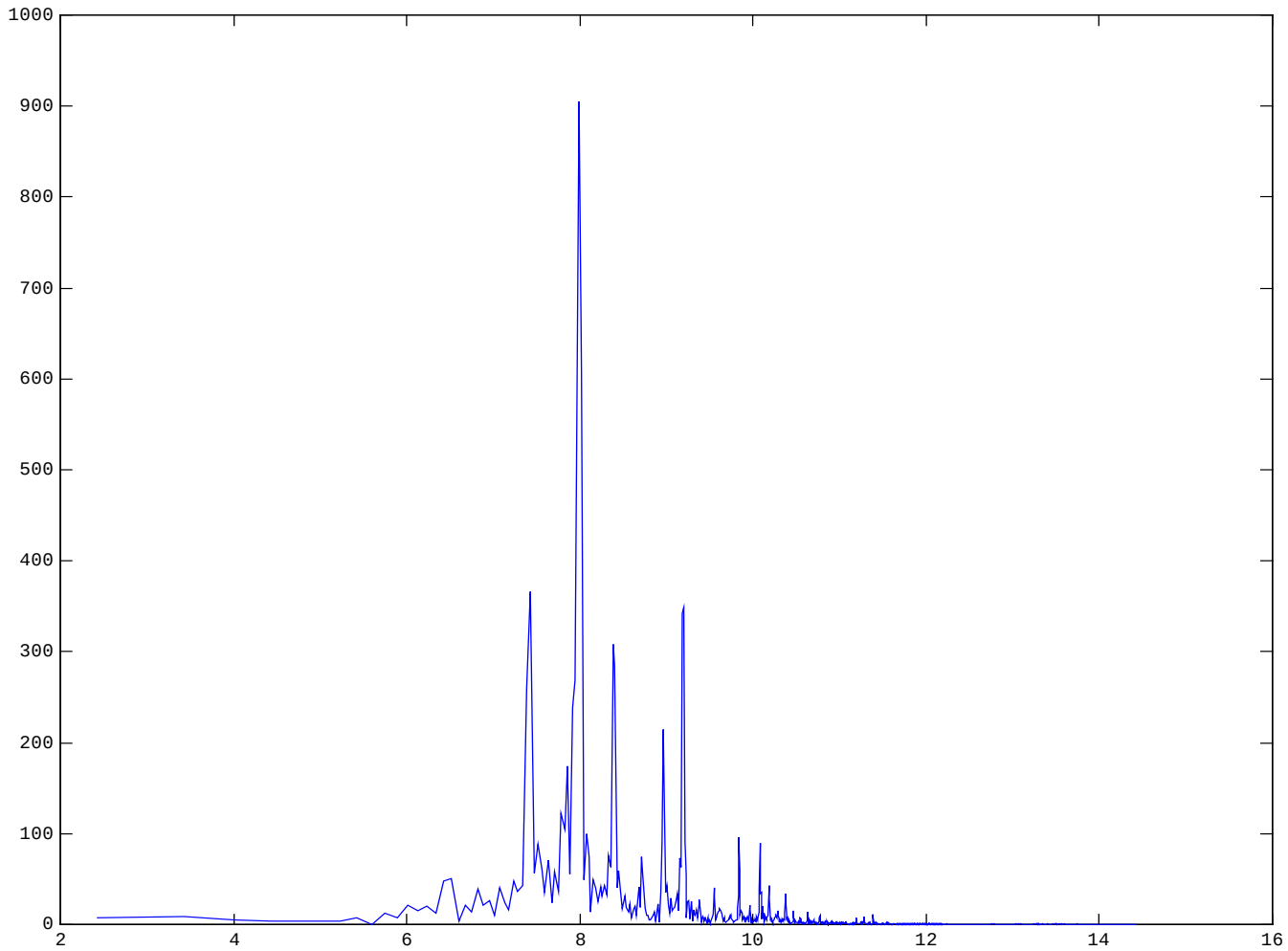


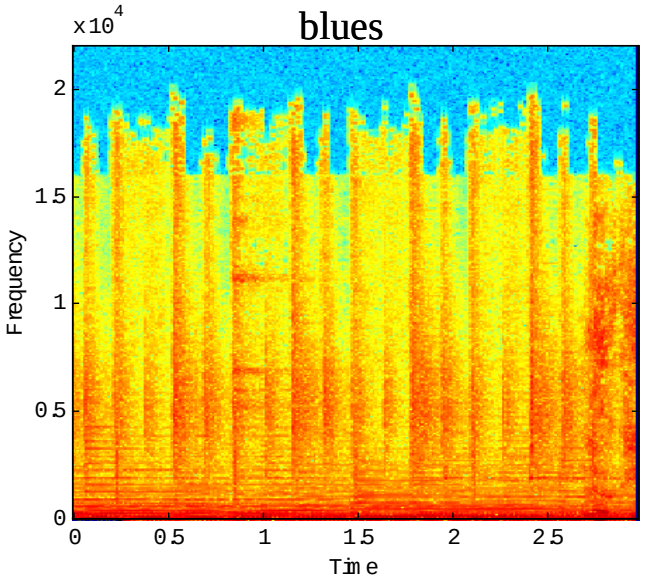


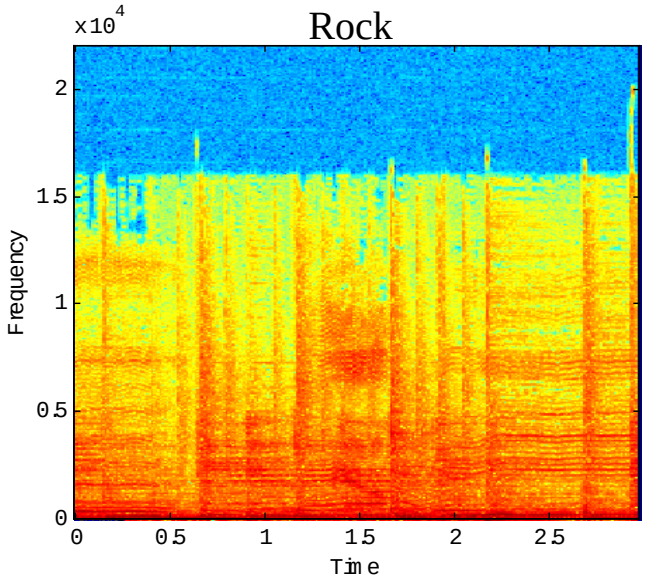
Spektrální analýza signálu

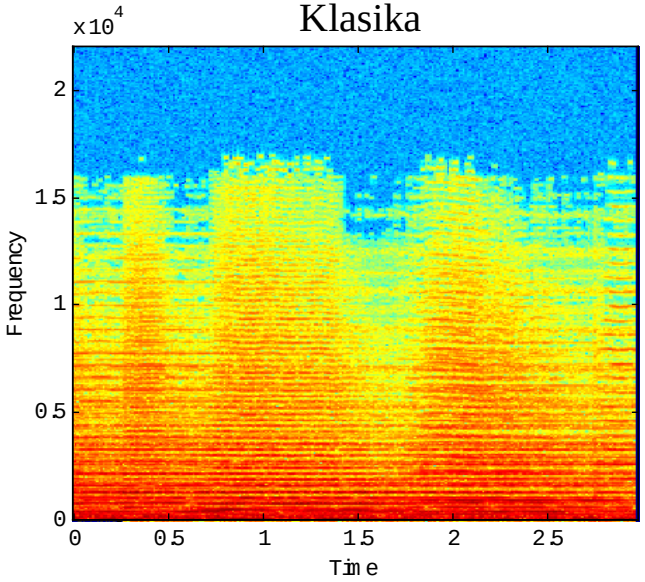
- WAV: časová posloupnost amplitud - vzorků
- Windowing: technický krok
- DFT (FFT): časová doména -> frekvenční doména
  - N vzorků -> N amplitud sinů + N amplitud cosinů
    - číslo vzorku ve FD značí poměr ke vzorkovací frekvenci:
      - » freq_vzorku = X/N * vzorkovací_freq
  - převod do polární reprezentace (magnituda + fáze):
    - A*cos(x) + B*sin(x) = M*cos(x+fi)
    - fázi ignoruji
  - Nyquistova věta: vzorkovací frekvence F -> nejvyšší frekvence v signálu F/2
    - potřebuji jen spodních N/2 vzorků
  - příklad: 1024 vzorků -> FFT -> 512 frekvenčních koeficientů, rozlišení: 44100/2/512 = 43Hz

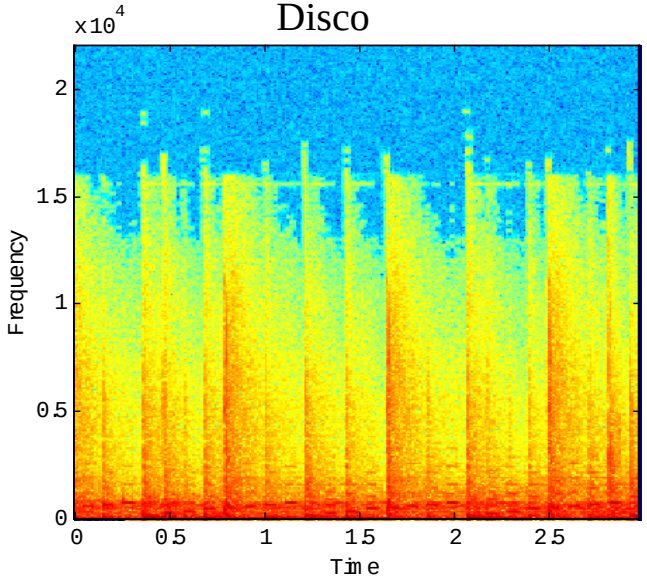
Spektrální analýza signálu




blues
x10^4
Frequency
Time
0
0.5
1
1.5
2
2.5


Rock
x10^4
Frequency
Time
0
0.5
1
1.5
2
2.5


Klasika
x10^4
Frequency
Time
0
0.5
1
1.5
2
2.5


Disco
x10^4
Frequency
Time
0
0.5
1
1.5
2
2.5

Spektrální analýza signálu

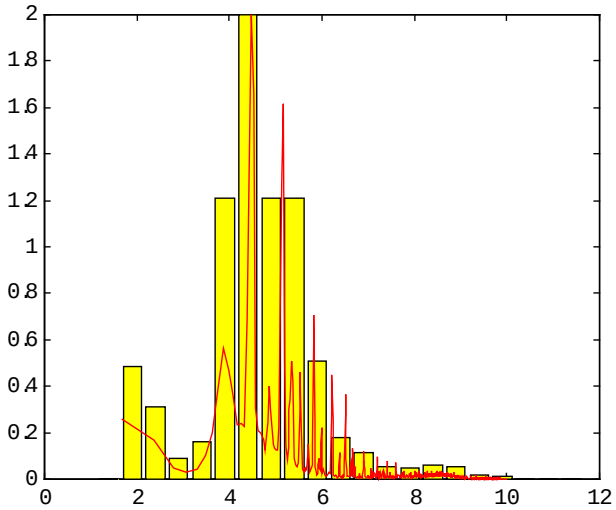
- problém: jediný segment obsahuje příliš mnoho šumu a „náhodných“ frekvencí
  - ne: větší okno (zjemní analýzu, ale zachová šum)
  - ano: průměr několika segmentů (zvýrazní frekvence vyskytující se ve všech vzorcích)
  - ano: průměr několika sousedních frekvencí (bude dále)

Spektrální analýza signálu

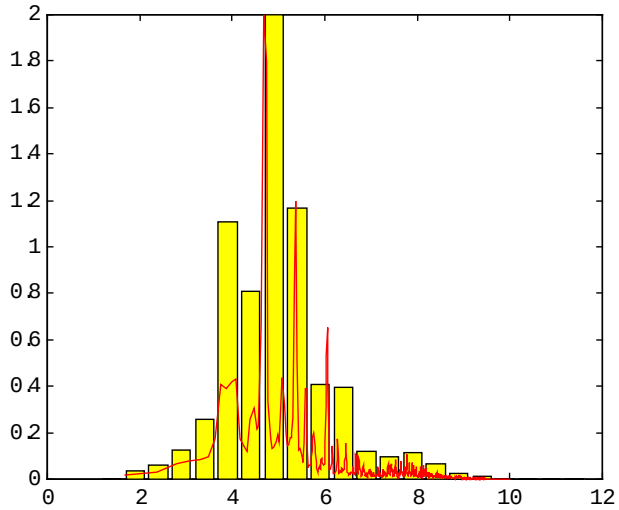
- problém: vektor segmentu je na NS příliš dlouhý
  - sdružení určitého intervalu frekvencí do jedné hodnoty
    - sčítáním se zároveň dále odstraňuje šum a fluktuace
  - volba intervalů („kapes“)
    - lineární (např. 1-500Hz, 500-1000Hz,...)
      - neodpovídá modelu ucha, vůbec nevystihuje nižší frekvence, neosvědčilo se
    - logaritmická (např. 256-512Hz,512-1024Hz)
      - odpovídá modelu ucha, nejlepší výsledky
    - oktávová (8 intervalů: 27.5(A0)...7040Hz(A8))
      - podstatou téměř shodná s logaritmickou, podobné výsledky
    - tónová (12 nespojitých intervalů)
      - jeden interval odpovídá všem tónům, nezávisle na oktávě (tóny zní stejně) -> invariance vůči transpozici
      - průměrné výsledky
    - „všetónová“ (98 intervalů pro každou notu)
      - motivace: každý styl má své oblíbené frekvence
      - výsledek: charakter náhodného šumu ;-)

Spektrální analýza signálu

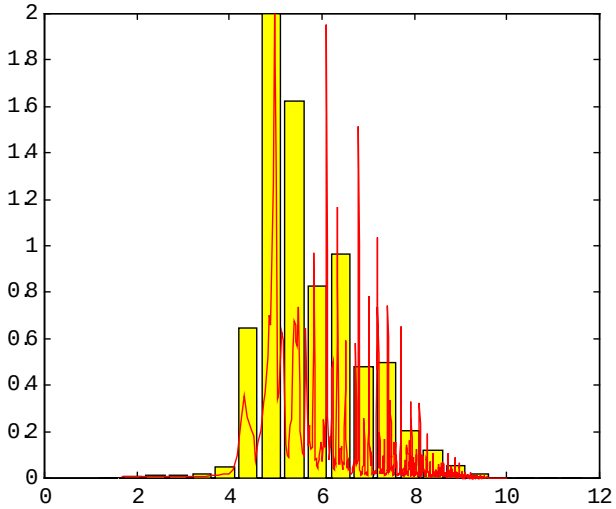
blues



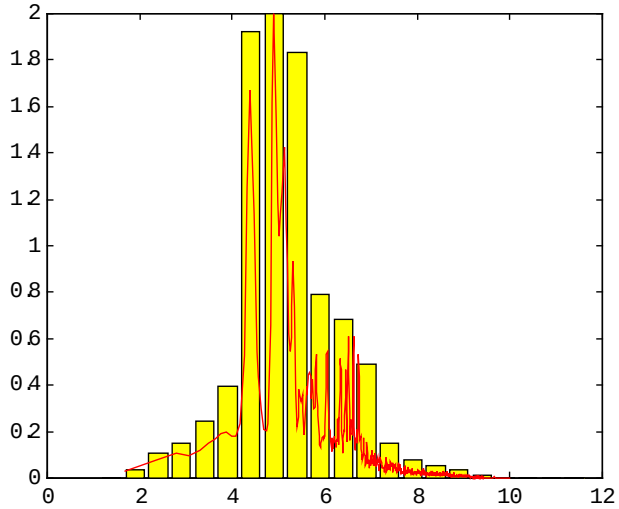
Rock



Klasika



Disco



Sběr dat

- 5 kategorií, 70 skladeb, 4,5 hodin muziky, 2.8 GB dat:
  - jazz (3 interpreti)
  - klasika (3 interpreti)
  - rock (4 interpreti)
  - folk (3 interpreti)
  - disco (2 interpreti)
- vstup: hlasitostně normalizované WAV soubory
  - 44.1kHz, stereo
- výstup: 5330 vzorků
  - vynechání 15s ze začátku a konce skladby (nereprezentativní)
  - vzorky dlouhé 10s, extrahované vždy po 3s

Analýza dat

- Nutná extrakce rysů vzorku
  - délka nezpracovaného vstupního vektoru při vzorkovací frekvenci 44.1kHz a 10s vzorku je 441 000
- Frekvenční doména
  - vhodná pro zpracování
  - detailní popis krátkého úseku (charakteristika)
  - statistický popis dlouhého (dynamika)

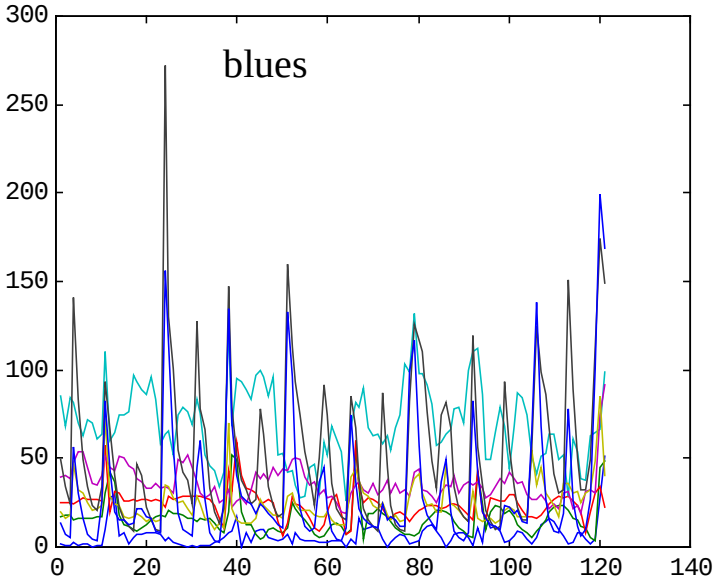
Analýza dat - charakteristika

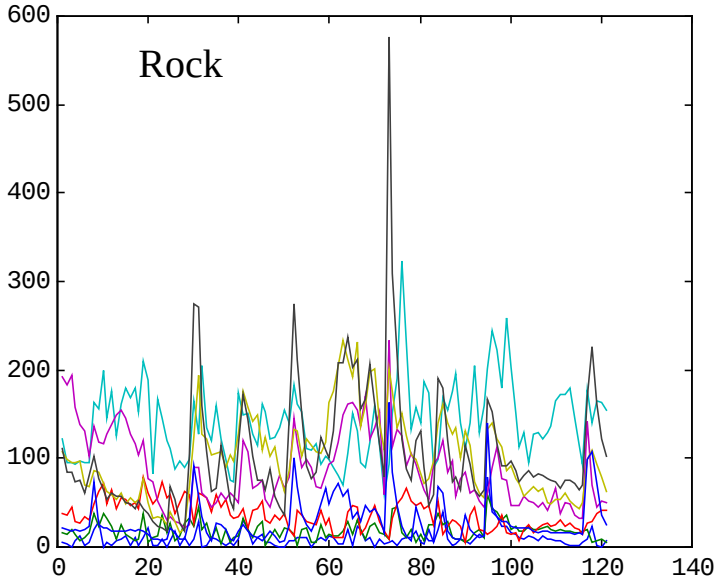
- provedení spektrální analýzy vzorku
  - 10x po 100ms s polovičním překryvem sousedů (celkem: 1s)
    - délka okna: 8192 vzorků (cca 200ms) -> detekuje 4096 frekvencí
  - volba logaritmického rozložení intervalů (16 kapes)
    - dávalo ze všech rozložení nejlepší výsledky, ačkoliv rozdíl nebyl veliký
  - obdrželi jsme 16 kapes s 10 hodnotami

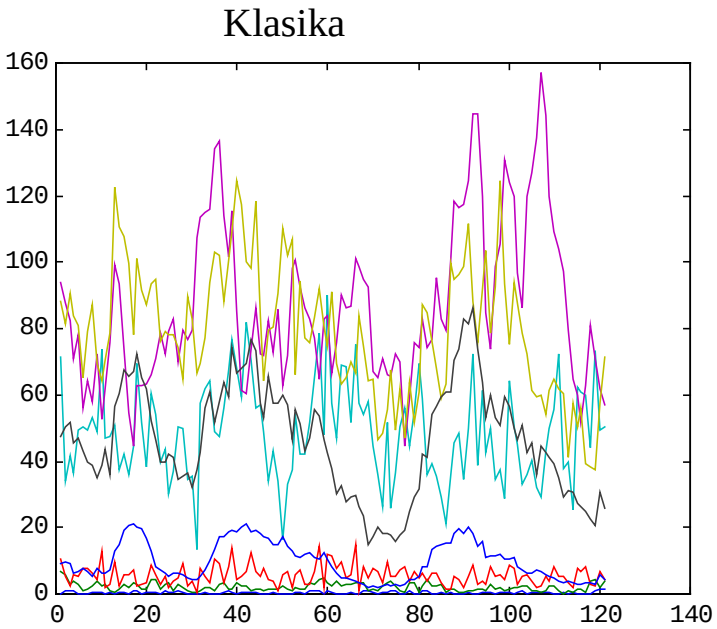
- zisk dat pro NS
  - součet všech 10 hodnot pro každou kapsu [16 hodnot]
    - nejlepší výsledky
    - sčítáním odstranění šumu a krátkodobých fluktuací
  - směrodatná odchylka pro každou kapsu [16 hodnot]
  - kvartily pro každou kapsu [48 hodnot]
    - oba rovněž dobré výsledky, ale nakonec nepoužity

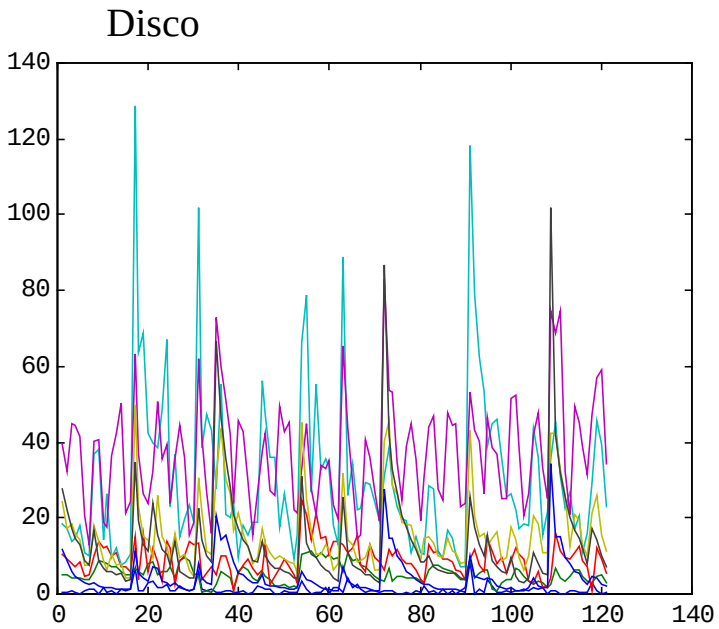
Analýza dat - dynamika

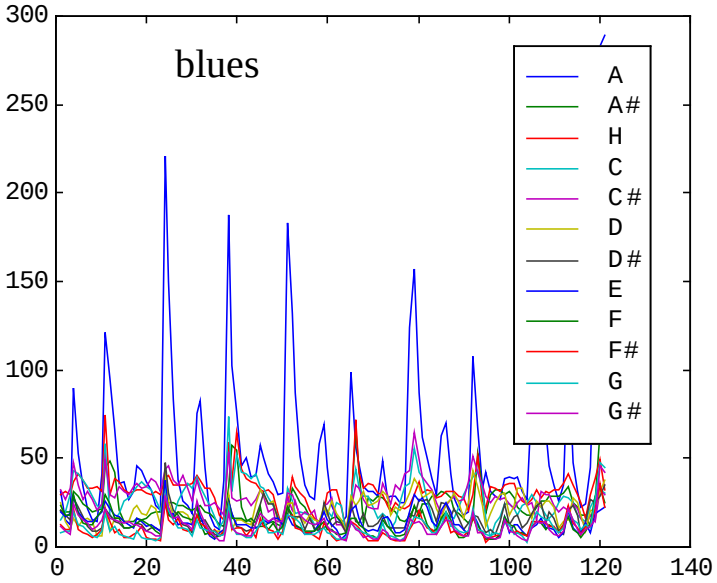
- provedení spektrální analýzy vzorku
  - 280x po 150ms (celkem: 7s)
    - délka okna: 1024 vzorků (cca 30ms) -> detekuje 512 frekvencí
  - volba logaritmického rozložení intervalů (16 kapes)
    - opět dávalo ze všech rozložení nejlepší výsledky
    - idea: vzít dynamiku rozloženou logaritmicky a charakteristiku např. podle tónů, čímž poskytnu více dat (v praxi bohužel nezlepšilo výsledek klasifikace)
  - obdrželi jsme 16 kapes s 280 hodnotami
- zisk dat pro NS
  - boxplot (kvartily a min/max bez odlehlých pozorování) pro každou kapsu [80 hodnot]
    - výborné výsledky
  - „výraznost“
    - = max z hodnot kapsy / průměr z hodnot kapsy
    - z toho průměr, medián, odchylka [3 hodnoty]

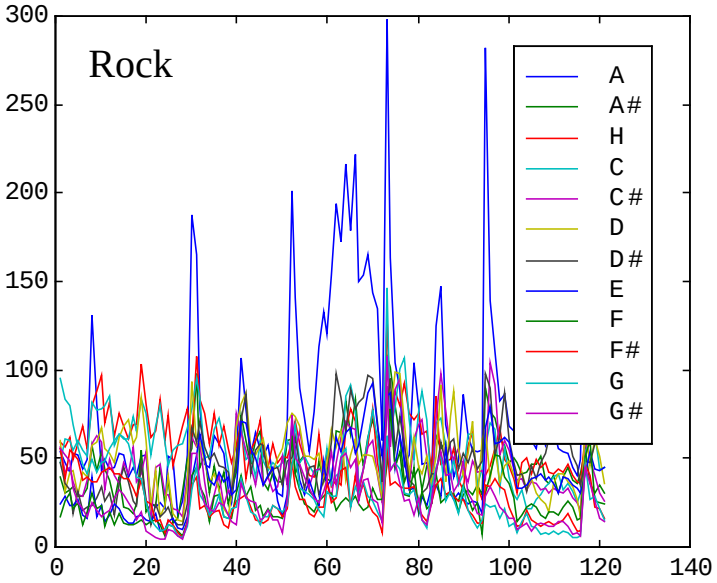

blues
300
250
200
150
100
50
0
0
20
40
60
80
100
120
140

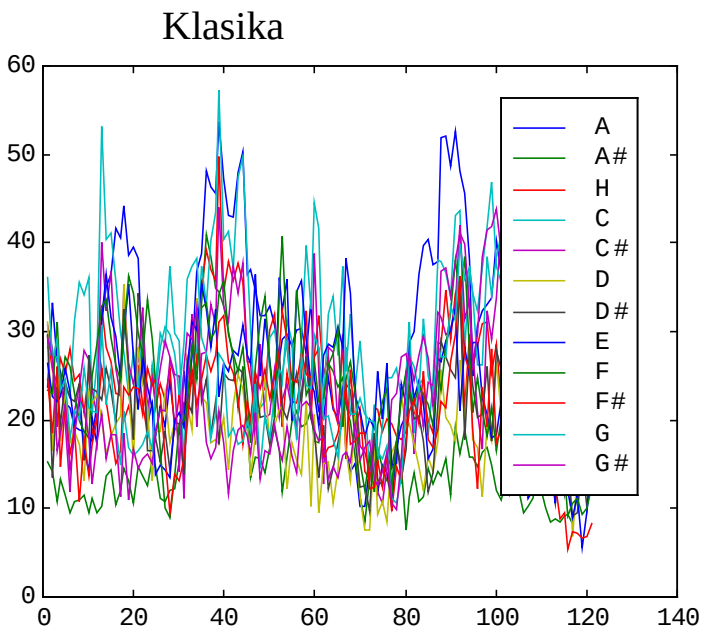

Rock
600
500
400
300
200
100
0
0
20
40
60
80
100
120
140

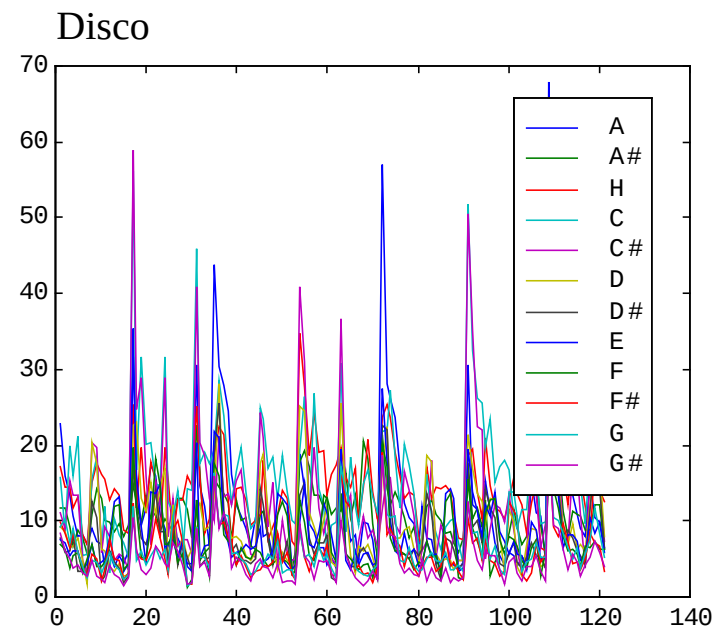

Klasika
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140

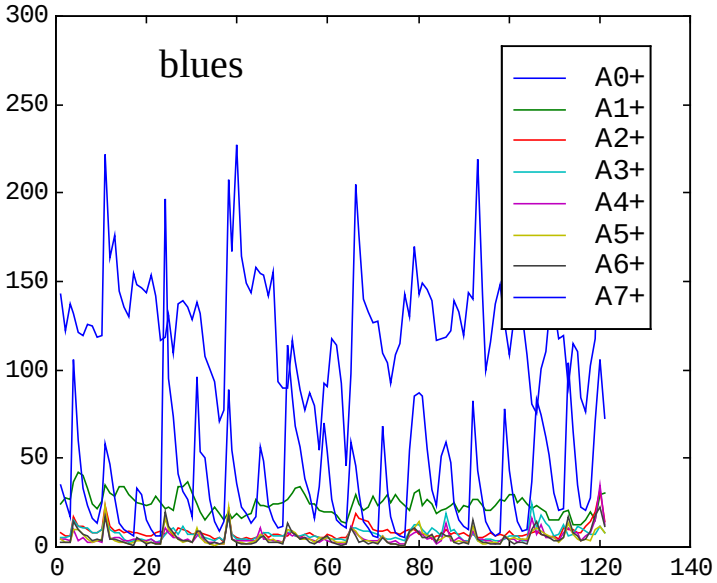

Disco
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140

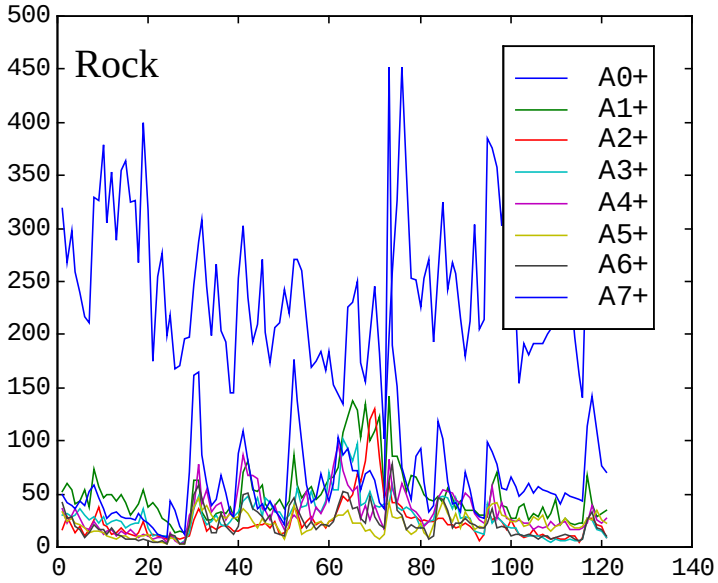

blues
A
A#
H
C
C#
D
D#
E
F
F#
G
G#
0
50
100
150
200
250
300
0
20
40
60
80
100
120
140

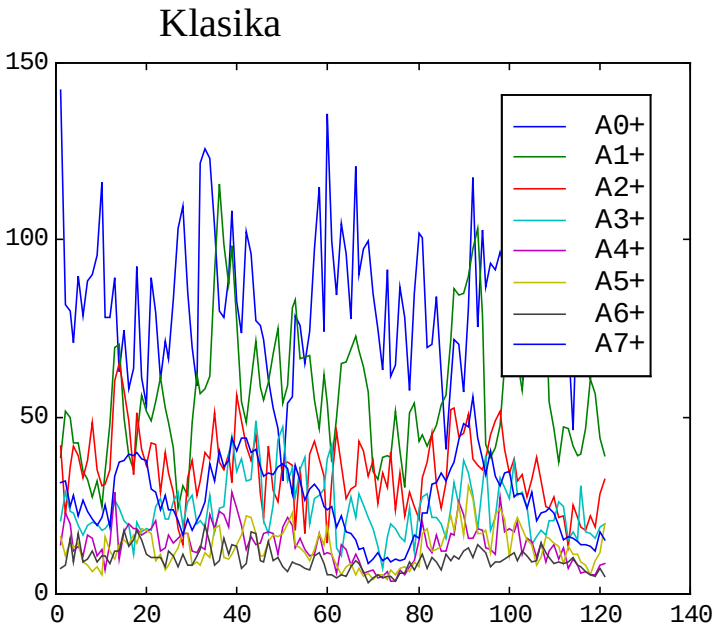

Rock
A
A#
H
C
C#
D
D#
E
F
F#
G
G#
0
50
100
150
200
250
300
0
20
40
60
80
100
120
140

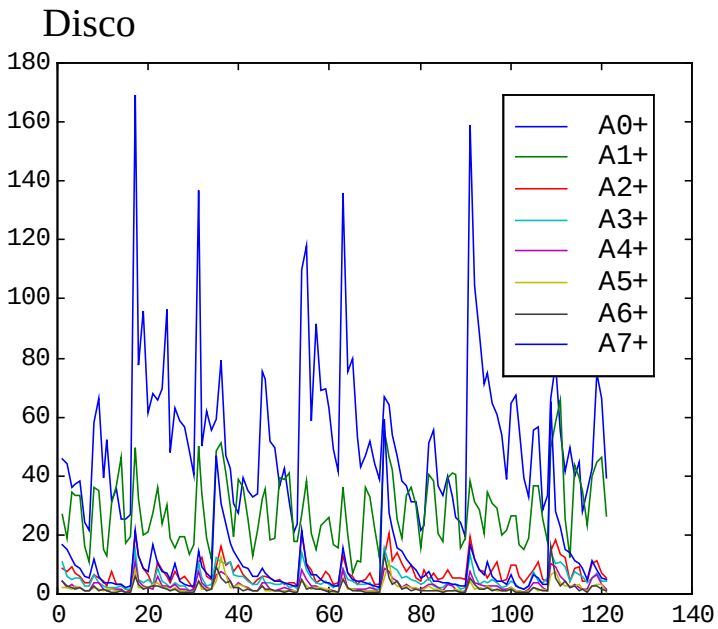

Klasika
A
A#
H
C
C#
D
D#
E
F
F#
G
G#
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140


Disco
A
A#
H
C
C#
D
D#
E
F
F#
G
G#
0
10
20
30
40
50
60
70
0
20
40
60
80
100
120
140


blues
A0+
A1+
A2+
A3+
A4+
A5+
A6+
A7+
0
50
100
150
200
250
300
0
20
40
60
80
100
120
140


Rock
A0+
A1+
A2+
A3+
A4+
A5+
A6+
A7+
0
50
100
150
200
250
300
350
400
450
500
0
20
40
60
80
100
120
140


Klasika
A0+
A1+
A2+
A3+
A4+
A5+
A6+
A7+
0
50
100
150
0
20
40
60
80
100
120
140


Disco
A0+
A1+
A2+
A3+
A4+
A5+
A6+
A7+
0
20
40
60
80
100
120
140
160
180
0
20
40
60
80
100
120
140

Analýza dat - dynamika

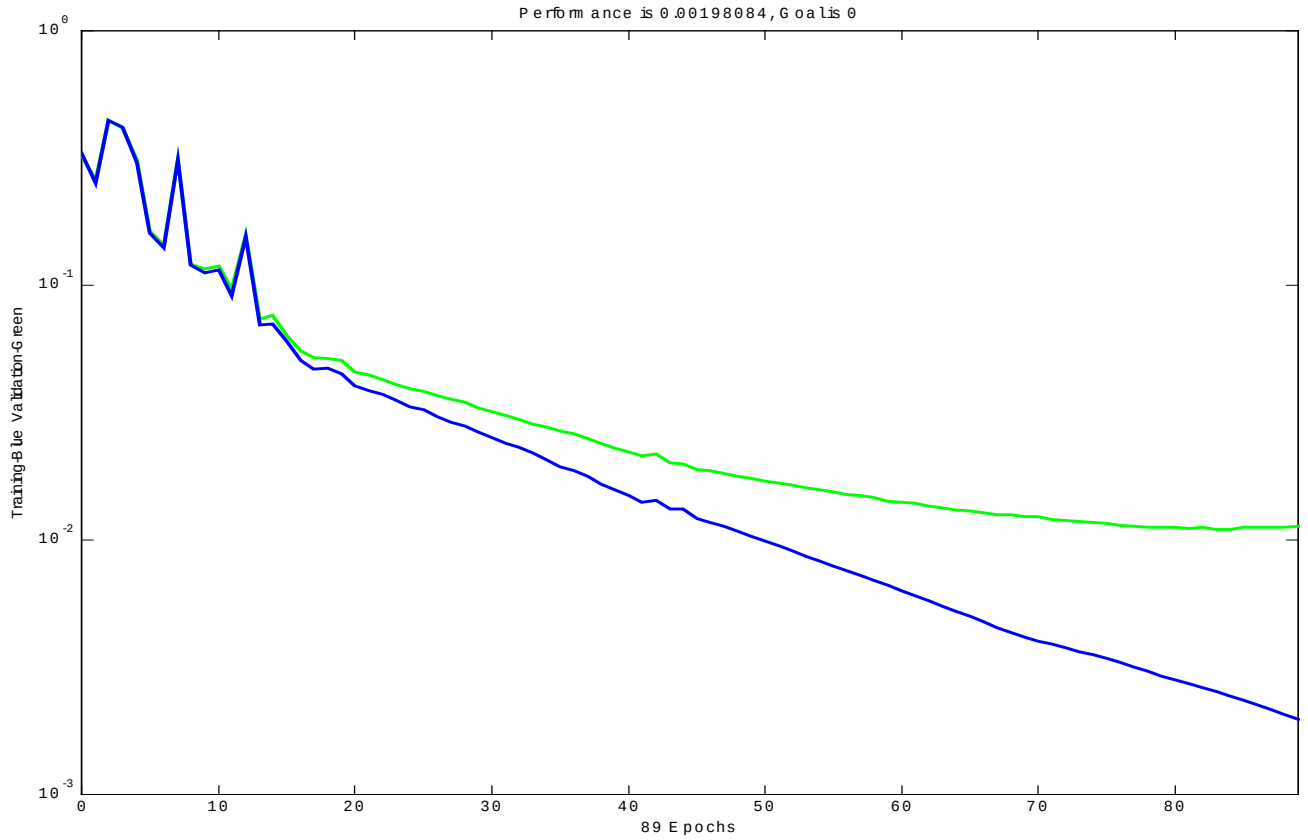
- další algoritmy (Matlabové skripty z Internetu)
  - detekce beatu
    - spolehlivé, používám průměrnou délku doby a její rozptyl
  - Linear Predictive Coding, Mel Frequency Cepstral Coefs
    - algoritmy pro charakteristiku mluvené řeči
    - relativně dobré výsledky, ale jejich přidáním k finálnímu vektoru kupodivu úspěšnost nestoupla
- charakteristika beatu
  - 10s vstupní vzorek analýzy rozdělen na úseky podle beatů
    - vezmu prvních 9 celých (tj. nejméně 2 celé takty)
  - každý úsek zanalyzován samostatně
    - logaritmické rozdělení na 12 kapes, pro každou část tedy 12 hodnot
  - z výsledných 9x12 hodnot opět spočten boxplot [60 hodnot]
  - velmi dobré výsledky

Analýza dat - shrnutí

- celková délka vstupního vektoru: 161
  - 16: charakteristika přes součty
  - 80: boxplot dynamiky
  - 2 : beat
  - 3 : výraznost
  - 60: charakteristika beatu
- PCA analýza vede ke zhoršení výsledků
- Vzorky rozděleny na učící, validační a testovací
  - testovací sada se skládá z 10 písniček
    - žádný vzorek z těchto skladeb se nevyskytl v učící množině

Učení

- BP síť (trainrp)
- výstup sítě: uspořádaná pětice hodnot 0..1
  - maximální složka je brána jako číslo třídy
- nejvhodnější architektura: [161, 240, 5]



Výsledky

ZAŘAZEN⇒ PATŘÍ ⇓	DISCO	FOLK	JAZZ	KLASIKA	ROCK
DISCO	166	8	1	2	3
FOLK	1	185	7	6	1
JAZZ	0	2	345	7	3
KLASIKA	0	1	0	466	0
ROCK	2	2	0	0	376

- úspěšnost: 97%
- data jsou z validační množiny
  - síť již danou skladbu viděla

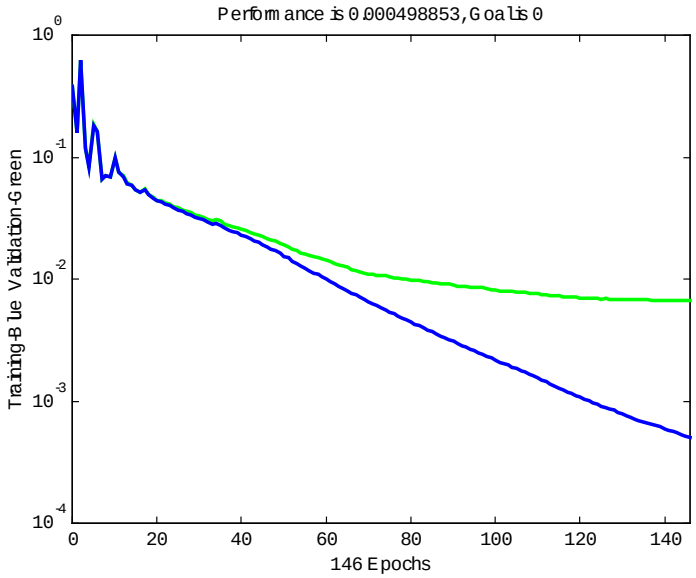
Výsledky

ZAŘAZEN⇒ PATŘÍ ⇓	DISCO	FOLK	JAZZ	KLASIKA	ROCK
DISCO	113	15	1	5	2
FOLK	0	44	27	8	13
JAZZ	0	1	55	0	2
KLASIKA	7	1	7	70	0
ROCK	5	7	2	12	183

- úspěšnost: 80%
- data jsou z testovací množiny
  - síť během učení danou skladbu neviděla

Výsledky

- Klasifikace interpretů
  - stejné podmínky, stejná NS
  - výstupní vektor: 15 složek pro 15 interpretů



- Úspěšnost
  - 94% pro data z validační množiny
  - pouze 11% pro data z testovací množiny
    - jen 2x lepší než náhodná volba
    - dáno malým množstvím vzorků jednoho interpreta (cca 3-4x menší množství než na jeden hudební styl)

Závěr

- Klasifikace byla vcelku úspěšná
  - vhodná data: relativně ortogonální výběr stylů a skladeb
  - záleží zejm. na množství zpracovaných dat, než na jejich přesném výběru a nastavení učícího algoritmu
- Doporučení:
  - The DSP Guide (www.dspguide.com)